



2015

Text Analytic System: Document Similarity

Ryan Murphy

Virginia Commonwealth University

James Cecil

Virginia Commonwealth University

Joseph Contarino

Virginia Commonwealth University

Follow this and additional works at: <https://scholarscompass.vcu.edu/capstone>

 Part of the [Computer Engineering Commons](#)

© The Author(s)

Downloaded from

<https://scholarscompass.vcu.edu/capstone/10>

This Poster is brought to you for free and open access by the College of Engineering at VCU Scholars Compass. It has been accepted for inclusion in Capstone Design Expo Posters by an authorized administrator of VCU Scholars Compass. For more information, please contact libcompass@vcu.edu.



Text Analytic System

Document Similarity



Abstract

- Text analytics is a critical function to knowledge discovery.
- Our algorithm processes web-based texts embedded in HTML pages and analyzes them to determine similarity.
- By analyzing the similarity of these HTML documents, we are helping the Idaho National Laboratory to keep redundant data out of the database. Without proper parsing of similar data, repetitive entries may clog the system with unneeded information.

Natural Language Processing

Term-Document Matrix: A method of analyzing the frequency of terms among documents. Rows correspond to documents, and columns correspond to terms.

Consider the sentences "I like apples" and "I like oranges".

	I	like	apples	oranges
Sentence 1	1	1	1	0
Sentence 2	1	1	0	1

Cosine Similarity: One way of describing how similar two documents are is to treat their rows in the term-document matrix as vectors, and then calculate the angle between them.

$$\text{Cos}() = \frac{\text{Vector1} \cdot \text{Vector2}}{|\text{Vector1}| |\text{Vector2}|}$$

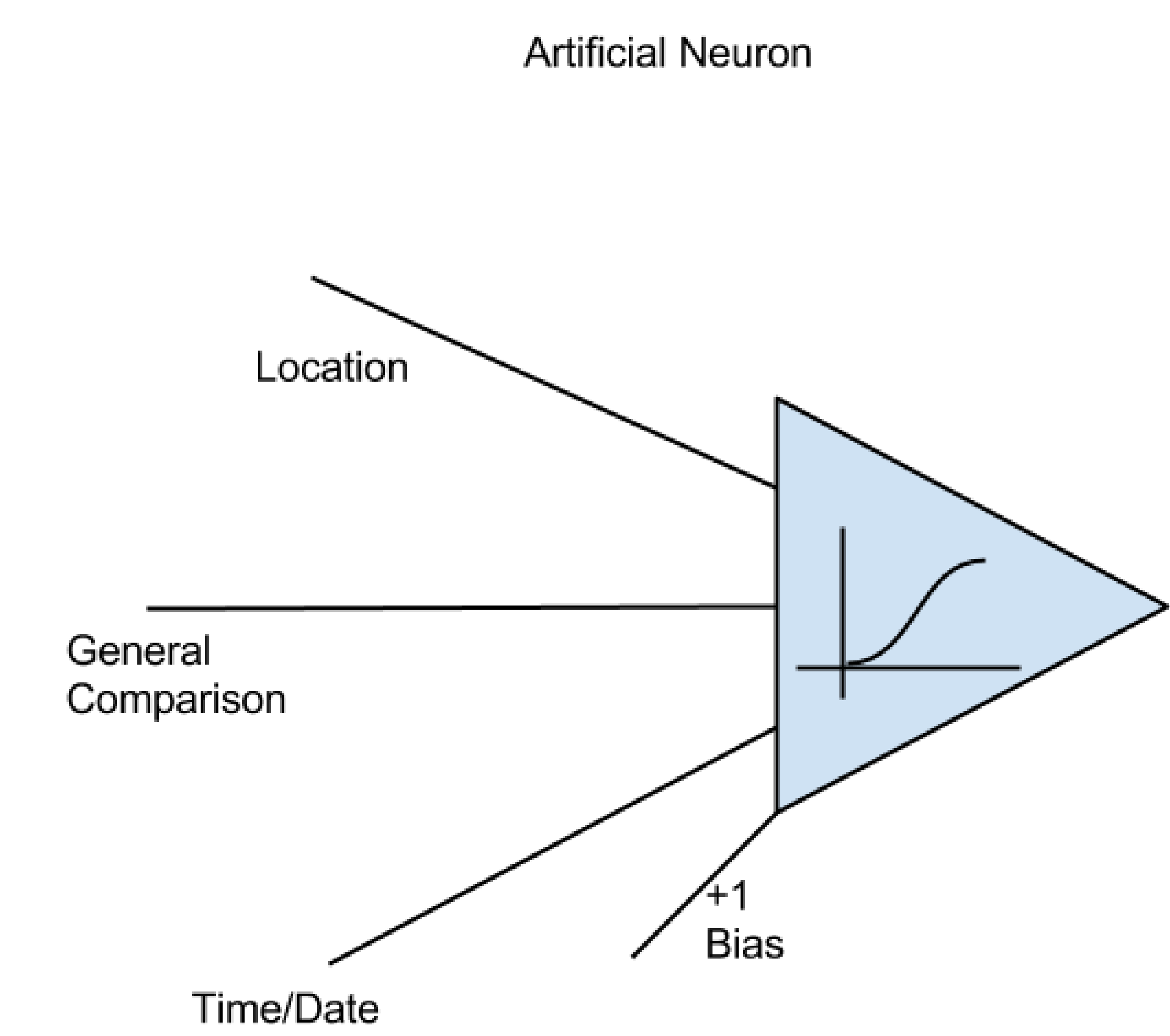
The result of the above matrix is:

$$\text{Cos}() = \frac{1 * 1 + 1 * 1 + 1 * 0 + 0 * 1}{\sqrt{1^2 + 1^2 + 1^2 + 0^2} * \sqrt{1^2 + 1^2 + 0^2 + 1^2}}$$



Results

- Results from cosine similarity analysis fed into an artificial neuron, which separates data into two fields: similar and dissimilar



- The neuron uses a soft activation function, which biases the output naturally towards either 0.0 or 1.0, representing dissimilar and similar comparisons, respectively

Overview

